comis

An R Package to Read CCCCO MIS Files

Christian Million
Data Analyst
Yosemite Community College District

Goals

Main Goal

- Showcase benefits of developing internal packages with R.

Along the way...

- Inspire!
- Use `comis` as a motivating example
- Why is package development worthwhile?
- Learn **too** much about MIS files.

```r
library(comis)

# Read Referential File
CB <- read_ref("path/to/CB223.txt")

# Read Submission File
XB <- read_sub("path/to/U59217XB.DAT")
```

What is `comis`?

An internally developed R package

Purpose

Read and Format:

- MIS **Submission** Files
- MIS **Referential** Files

MIS 101

Submission Files: Districts submit MIS data to state via these files

- ~ 25 files | 396 elements

Referential Files: Districts retrieve these files from Data-On-Demand

- ~ 27 files | 406 elements

The Challenge

- MIS Data is important

- We want to easily use it / analyze it

- **The Challenge:** Reading the MIS data into R is difficult and error prone

Submission Files - Challenges

- Fixed Width Format

- No Column Names

- Numbers that should be characters / dates

- Missing values (NA)
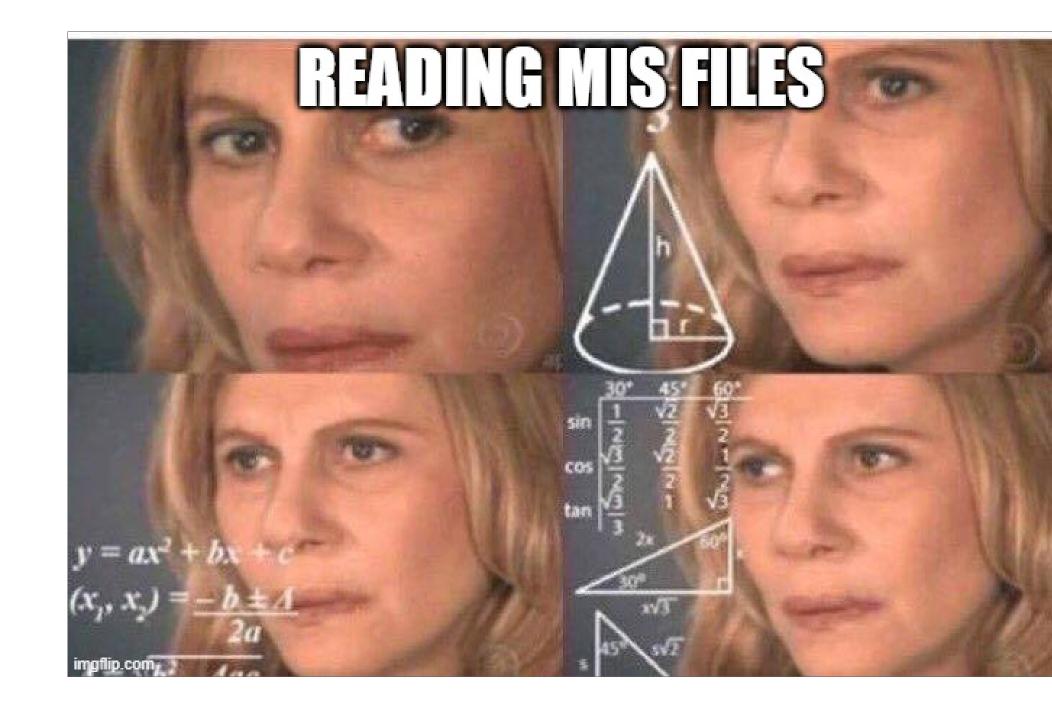
- Trailing white space

- Implied decimal points

Referential Files - Challenges

- Tab Delimited :)

- No Column Names

- Numbers that should be characters / dates

- Missing values (NA)

- Trailing white space

- ~~Implied decimal points~~

- *Different date format than submission*

```r
library(comis)

# Column names are DED Codes.
# like "GI01", "CB00", "CB01"
read_ref_repo("CB", "217")

# Column names are words.
# like "COLLEGE_ID", "COURSE_ID", "CONTROL_N
read_ref_repo("CB", "217", desc = TRUE)
```
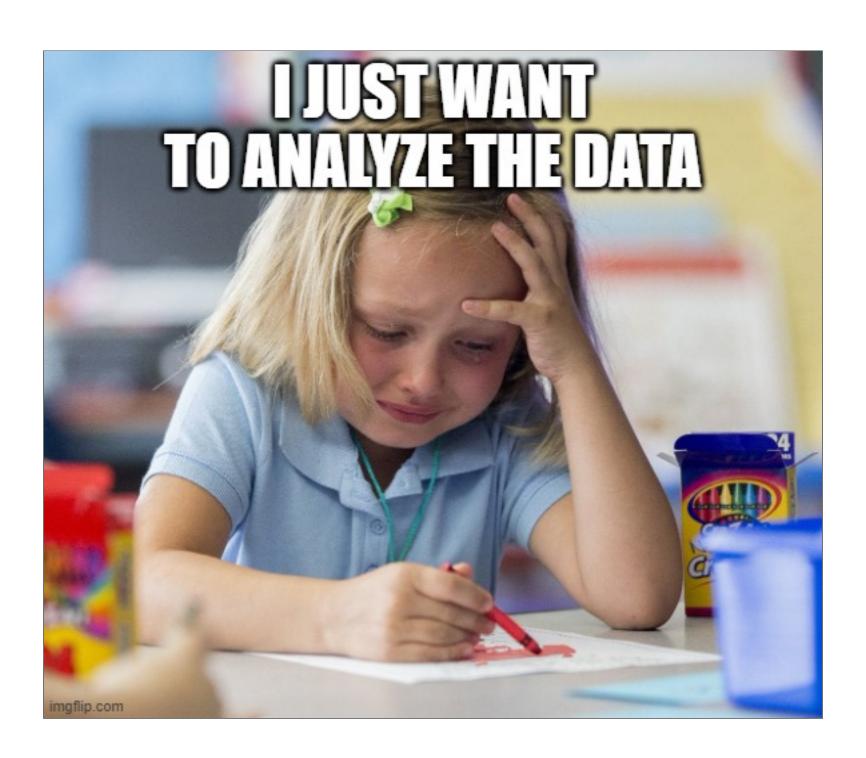
READING MIS FILES

Yikes

Imagine writing code to handle this for each analysis:

- A lot to re-remember

- Cognitively taxing to implement

- Takes time

- Updates to multiple scripts

- Copy / paste errors

- Makes scripts more difficult to read

- Unfulfilling

- Lots of overhead before analysis can begin

Before `comis`

```r
# Load Libraries -----
library(dplyr)
library(readr)


# Define Names, Types, and Widths -----
CB_col_names <- c('GI90', 'GI01','GI03', paste0("CB0",0:9), paste0("CB",10:2
CB_col_types <- rep("c", length(CB_col_names))
CB_col_width <- CB <- c(2,3,3,12,12,68,6,1,1,length(109:112),length(113:116)

XB_col_names <- c('GI90', 'GI01', 'GI03', 'GI02', 'CB01', paste0('XB0',0:9),
XB_col_types <- rep("c", length(XB_col_names))
XB_col_width <- c(2,3,3,3,12,6,1,6,6,1,length(44:47), length(48:51),1,1,1,1,

# Read the source data -----
CB_src <- readr::read_tsv("path/to/U59223CB.DAT",
                          col_names = CB_col_names,
                          col_types = CB_col_types,
                          trim_ws = TRUE)
```

I JUST WANT
TO ANALYZE THE DATA

After `comis`

```r
# Load Libraries -----
library(comis)

# Read Data -----
CB <- read_sub("path/to/U59223CB.DAT")
XB <- read_sub("path/to/U59223XB.DAT")
```

Additional Features

- Contains useful data found on CCCCO websites

- Read many files at once

- Read from repo

- Use DED Name or Descriptive Name

read_ref {comis}

Read an MIS Referential File

Description

Read in an MIS Referential File as a data.frame.

Usage

```
read_ref(file, domain = NULL, fmt = TRUE, desc = FALSE, trim_ws = TRUE)

read_ref_repo(
  domain,
  term = NULL,
  repo = getOption("comis.repo.referential"),
  district = getOption("comis.district"),
  ...
)
```

Arguments

`file`	chr A path including the filename and extension.
`domain`	chr The domain of the file (e.g., "CB", "SX")
`fmt`	lgl format the columns according to their data element format
`desc`	lgl whether to use the data entry description for column names
`trim_ws`	lgl trim leading / trailing whitespace of each column
`term`	char 3 digit term identifier. If `NULL` it will use all available terms for a given domain.
`repo`	char a path to a directory containing referential files.
`district`	the 3 digit district code
`...`	additional parameters passed to `read_ref`

Details

The referential file for a specific domain comes from the CCCCO Data-on-Demand site, and is a general, summarized versions of this data are what populate sites like DataMart, Student Succes

You should be able to read multiple files at once as long as all of the files exist and correspond t

Benefits of `comis`

- Easier to tell what's happening

- Reduces cognitive overhead

- Get to analysis faster and with more confidence

- Documentation contained within the package

- Updates made in one spot (instead of throughout various scripts)

- **Shifts focus to what's important - Using the Data**

Why Develop Internal R Packages?

- Addresses problems specific to the institution

- Reasonable defaults

- Abstracts common tasks

- Maintainable

- Easily share code with others

- Business logic is located in one place

Recommended Reading: Building a Team of Internal R Packages

Thanks!

Contact

Christian Million

Data Analyst

Yosemite Community College District

millienc@yosemite.edu

Slides were created with **Quarto**

Slide Deck: **https://christian-million.github.io/comis_presentation/**

Presentation source code: **https://github.com/christian-million/comis_presentation**